ÖREBRO UNIVERSITET
AKADEMIN FÖR
NATURVETENSKAP OCH TEKNIK

Lösningar till tentamen i

Numeriska metoder för civilingenjörer

DT508G

2018-11-02

Anmärkning: Vid rättning betyder ✓ "rätt", f "fel", (✓)"rätt efter fel". Observera att lösningsförslag inte är en fullständig lösning.

1. Lös följande delproblem. [10p]

   (a) Gör en flyttalsberäkning i dubbel precision av $2.3 - 2$ dvs representera först 2,2.3 som flyttal och gör sedan subtraktionen med avrundning till närmaste. [4p]

   (b) Beräkna det relativa felet som uppkommer vid beräkningen i (a) och visa att det är mindre än halva maskinnogrannheten. [2p]

   (c) Gör en uppskattning av beräkningsfelen som uppkommer då man med flyttal beräknar en summa av n tal dvs $\sum x_k$. [4p]
   Ledning; Det gäller för alla elementär beräkningar att $fl(x \otimes y) \leq (1 + \epsilon), |\epsilon| \leq \epsilon_M$ där ϵ_M är maskinnogrannheten. Använd denna relation på multiplikationerna och additionerna i inre produkten och finn sedan en övre begränsning av det relativa felet.

   Lösning:

(a),(b) Vi har $fl(2.3) = 1.10\overline{1001} \times 2$ och $fl(2) = 1.00\ldots0 \times 2$ där mantissan har 52 bitar vilket ger $fl(2.3) = 1.10\ldots10 \times 2$ och $fl(fl(2.3) - fl(2)) = 1.010010110\ldots100110$ med 52 bitar i mantissan. Det sker ingen avrundning så det fel som uppkommer är 53e biten och bakåt i 0.3 dvs $0.\overline{1001} \times 2^{-52} \times 2^{-1} \leq 2^{-53} = \epsilon_M/2$.

(c) Vi har

$$fl\left[\sum_{k=1}^{n} fl(x_k)\right] = fl\left[\sum_{k=1}^{n} x_k(1+\epsilon_k)\right]$$

som ger om vi beräknar summan från vänster till höger (term 1 och 2 adderas först som sedan adderas till term 3 som adderas till term 4...)

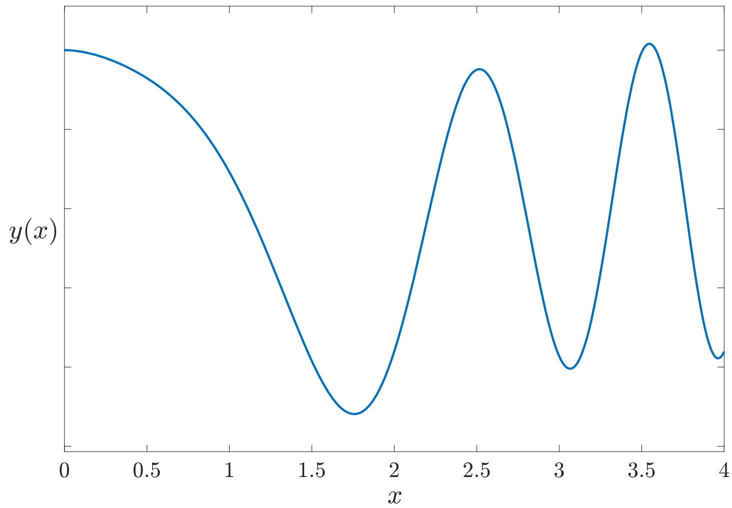
$$fl\left[\textstyle\sum_{i=1}^{n} fl(x_k)\right] = [[\cdots[x_1(1+\epsilon_1) + x_2(1+\epsilon_2)](1+\epsilon_{n+1}) + \\ + x_3(1+\epsilon_3)](1+\epsilon_{n+2}) + \ldots + x_n(1+\epsilon_n)](1+\epsilon_{2n})$$

och genom att använda triangelolikheten att antal gånger fås

$$|fl\left[\sum_{k=1}^{n} fl(x_k)\right] - \sum x_k| \leq c\sum |x_k|\,|\epsilon_k| + \mathcal{O}(\epsilon_M^2) \leq d\epsilon_M n + \mathcal{O}(\epsilon_M^2)$$

för några positiva konstanter c,d oberoende av n. Man kan göra detta mer formellt genom induktion.

2. Företaget Hiking AB planerar ett företegsevenemang och behöver då veta hur krävande banan är som ges av kurvan $y(x) = \cos(x^2) - xe^{-x}\sin(x), x \in [0, 4]$ se figur nedan. [10p]



Dom anlitar en fysiker som kommer fram till följande energiformel

$$E = \int_0^4 y_+(x)\,dx$$

där y_+ är den del av kurvan som har positiv lutning ($y'(x) > 0$). Fysikern påstår (med rätta) att det finns ingen exakt lösning. Hjälp företaget att räkna ut integralen numeriskt genom att lösa följande deluppgifter.

(a) Beskriv i detalj en metod som räknar ut nollställena till derivatan i det givna intervallet. Ange metodens konvergenshastighet. [4p]

(b) Inför en diskretisering som tar hänsyn till att y_+ inte är definerad på ett slutet intervall. [2p]

(c) Beskriv en metod som löser integralen, dvs beräknar E, och ange metodens trunkeringsfel. [4p]

Lösning:

(a) T.ex. Newton-Raphsons metod som har kvadratisk konvergenshastighet, se kursboken för detaljer.

(b) Diskretiseringen måste delas in i delintervall där derivatan är positiv. Det ger två delintervall säg $[a,b]$,$[c,d]$. I varje av dessa delintervall införs en diskretisering $x_k, k = 1,\dots,n$ med x_1,x_n givet av ändpunkterna i de två intervallen.

(c) Trapetsmetoden kan användas på respektive intervall, för detaljer se boken. Om diskretiseringspunkterna väljs ekvidistant $x_{k+1} - x_k = h$ blir trunkeringsfelet av storleksordning $\mathcal{O}(h^2)$.

3. Avgör om följande påståenden är sanna eller falska och ge en kort motivering till ert svar. [10p]

(a) Runges fenomen uppkommer vid lösning av ekvationer med Newton-Raphsons metod. [2p]

(b) Intervallhalvering har linjär konvergenshastighet. [2p]

(c) Gausselimination är alltid en stabil metod, dvs ger alltid en lösning med små relativa fel. [2p]

(d) Trapetsmetoden är en explicit metod för att lösa ickelinjära ekvationer. [2p]

(e) Sekantmetoden har konvergerar snabbare nära en lösning än fixpunktiteration. [2p]

Lösning:

(a) Nej, Runges fenomen uppkommer vid interpolation med högre ordningens polynom.

(b) Ja, eftersom felet alltid är mindre än intervallet som hela tiden halveras.

(c) Nej, inte utan att pivotera. Även med partiell pivotering är metoden i teorin inte stabil men i praktiken.

(d) Nej, metoden används antigen i samband med differentialekvationer eller för att lösa integraler numeriskt.

(e) Det beror på vilken fixpunktiteration som åsyftas. Newton-Raphson är en fixpunktiteration som har kvadratisk konvergenshastighet och är snabbare än sekantmetoden men andra fixpunktmetoder kan ha linjär konvergens som är långsammare.

4. Betrakta systemet [10p]

$$\begin{cases} y_1' &= t^2 + y_1 y_2 \\ y_2' &= y_1 + y_2 \end{cases}, \; t > 0,\; y_1(0) = y_2(0) = 1.$$

(a) Skriv om problemet på vektorform $\mathbf{y}' = \mathbf{f}(t,\mathbf{y})$ dvs definiera $\mathbf{f}$ och $\mathbf{y}$ samt begynnelsevillkoren på vektorform. [2p]

(b) Beskriv i detalj explicit Euler för detta problem i de givna vektorbeteckningar från (a). [4p]

(c) Ange noggrannhetsordningen på explicit Euler och föreslå en annan metod som är noggrannare. [2p]

(d) Om systemet är styvt ange en alternativ metod och dess noggrannhetsordning. [2p]

Lösning:

(a) Vi har $f_1 = t^2 + y_1 y_2$, $f_2 = y_1 + y_2$.

(b) Se boken kapitel 6.3.

(c) Noggrannhetsordningen är 1 dvs globala felet är av storleksordning h. En mer noggrann metod är Trapetsmetoden, se boken, som är av ordning 2. Den kräver dock en ytterligare funktionsberäkning i varje tidssteg samt att man löser ett ickelinjärt ekvationssytem i varje steg.

(d) Om systemet är styvt så kan man prova implicit Euler eller Trapetsmetoden som har ordning ett respektive två. [2p]

5. Givet matrisen [10p]

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 4 \\ 4 & 4 \\ 4 & 1 \end{bmatrix} \text{ och högerledet } \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

betrakta det linjära minstakvadratproblemet

$$\min_x \|b - Ax\|_2.$$

Lös följande deluppgifter.

(a) Sätt upp normalekvationerna antingen direkt eller genom att härleda dessa. Ni behöver inte lösa ekvationerna. [2p]

(b) Beskriv hur man kan lösa normalekvationerna med en Choleskyfaktorisering $A^T A = R^T R$. Ni behöver inte göra några detaljberäkningar utan endast ange delstegen utifrån faktoriseringen. [3p]

(c) Ange antalet beräkningar för metoden i (b) uttryckt i antal rader m och kolumner n i A. [1p]

(d) Antag nu att det beräknats en QR-uppdelning på formen [3p]

$$A = Q \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

Ange storlekarna på matriserna Q och R och hur man med dessa kan lösa minstakvadratproblemet.

(e) Ange antalet beräkningar för metoden i (d) uttryckt i antal rader m och kolumner n i A. [1p]

Lösning:

(a) Normalekvationerna blir $A^T Ax = A^T b$ där A är transponatet av A och $A^T A$ är en symmetrisk positivt definit matris. [2p]

(b) Bilda vektorn $c = A^T b$. Lös det undertriangulära systemet $R^T y = c$ och sedan det övertriangulära systemet $Rx = y$. [3p]

(c) Att bilda $A^T A$ kräver mn^2 additioner och multiplikationer samt Cholsekyfaktoriseringen $n^3/3$ beräkningar. [1p]

(d) Matrisen $Q \in \mathbb{R}^{m\times m}$ är en ortogonal matris och $R \in \mathbb{R}^{n\times n}$ är en övertriangulär matris. Vi får $\|b - Ax\|_2 = \|Q^T(b - Ax)\|_2 = \|c - [R; 0]x\|_2$ som med $c = [c_1; c_2]$ ger lösningen given av $Rx = c_1$. [3p]

(e) Den dominerande termen i antalet beräkningar är $2mn^2$ (både bildandet av c och lösandet av det övertriangulära systemet är av lägre ordning). [1p]

6. Betrakta problemet $Ax = \lambda x$ där $A \in \mathbb{R}^{n\times n}$ är symmetrisk och $x \in \mathbb{R}^n$,$\lambda \in \mathbb{R}$ är obekanta och lös följande uppgifter. [10p]

(a) Ange ytterligare en ekvation så att problemet får entydig lösning. [1p]

(b) Ge en algoritm för potensmetoden uttryckt i matrisen A och en approximation av egenvektorn x_k i iteration k. Algoritmen ska innehålla en approximation av egenvärdet. [2p]

(c) Visa att potensmetoden konvergerar mot egenvektorn motsvarande det till belopppet största egenvärdet. [4p]

(d) Beskriv en metod som konvergerar mot egenvektorn som hör till beloppet minsta egenvärdet. Formulera metoden så att inga onödiga beräkningar görs. [3p]

Lösning:

(a) T.ex. $\|x\|_2 = 1$ dvs Euklidiska längden av x lika med ett. [1p]

(b) Iterera $x_{k+1} = Ax_k, x_{k+1} = x_{k+1}/\|x_{k+1}\|_2$ (eller någon annan norm) där en approximation av till beloppet största egenvärdet är $\lambda = x_{k+1}^T x_k$. [2p]

(c) Antag att första approximationen är $x_0 = \sum_{j=1}^n c_j v_j$ där v_k är egenvektorerna till A (som är ortogonala eftersom A är symmetrisk). Efter k iterationer fås $x_{k+1} = Ax_k = A^k x_1 = \sum_j c_j \lambda_j^k v_j$ vilket ger att termen med det största egenvärdet kommer att dominera alltmer och slutligen med normalisering har vi $x_k \to v_m$ där λ_m är det till beloppet största egenvärdet. [4p]

(d) Gör en LU-faktorisering av A och lös för $k = 1,2,\ldots$ det linjära ekvationssystemet $Ax_{k+1} = x_k$ normalisera, $x_{k+1} = x_{k+1}/\|x_{k+1}\|_2$ genom att använda LU-faktoriseringen. [3p]